Journal of Siberian Federal University. Biology 3 (2015 8) 268-277
~ ~ ~

УДК 57:015 + 573.2

Seven-Cluster Structure of Larch Chloroplast Genome

Michael G. Sadovsky[a]*, Eugenia I. Bondar[b], Yulia A. Putintseva[b], Natalia V. Oreshkova[b,c], Eugene A. Vaganov[b] and Konstantin V. Krutovsky[b,d,e,f]

[a]Institute of Computational Modelling SB RAS
50/44 Akademgorodok, Krasnoyarsk, 660036, Russia
[b]Siberian Federal University
Genome Research and Education Centre
50a/2 Akademgorodok, Krasnoyarsk, 660036, Russia
[c]V. N. Sukachev Institute of Forest SB RAS
50/28 Akademgorodok, Krasnoyarsk, 660036, Russia
[d]Georg-August-University of Göttingen
2 Büsgenweg, Göttingen, D-37077, Germany
[e]N. I. Vavilov Institute of General Genetics RAS
3 Gubkin Str., Moscow, 119333, Russia
[f]Texas A&M University
HFSB 305, 2138 TAMU, College Station, Texas, 77843, USA

Received 20.01.2015, received in revised form 09.02.2015, accepted 26.04.2015

The paper presents a novel approach to study a nucleotide sequence structure with respect to the chloroplast genome DNA sequence analysis. A specific frequencies distribution pattern of the consecutive triple nucleotide fragments was identified in the chloroplast genome DNA sequence, which demonstrated a non-degenerated pattern with seven clusters.

Keywords: chloroplast genome, complexity, frequency dictionary, order, phase, triplet.

DOI: 10.17516/1997-1389-2015-8-3-268-277.

* Corresponding author E-mail address: msad@icm.krasn.ru

Семикластерная структура генома хлоропласта лиственницы

М.Г. Садовский[а], Е.И. Бондар[б], Ю.А. Путинцева[б], Н.В. Орешкова[б,в], Е.А. Ваганов[б], К.В. Крутовский[б,г,д,е]

[а]Институт вычислительного моделирования СО РАН
Россия, 660036, Красноярск, Академгородок, 50/44
[б]Сибирский федеральный университет
Научно-образовательный центр геномных исследований
Россия, 660036, Красноярск, Академгородок, 50а/2
[в]Институт леса им. В.Н. Сукачева СО РАН
Россия, 660036, Красноярск, Академгородок, 50/28
[г]Гёттингенский университет им. Георга-Августа
Германия, D-37077, Геттинген, ул. Бюсгенвег, 2
[д]Институт общей генетики им. Н.И. Вавилова СО РАН
Россия, 119991, Москва, ул. Губкина, 3
[е]Техасский агро-инженерный университет
США, HFSB 305, 2138 TAMU, штат Техас 77843, г. Колледж Стейшн

Проанализированы структуры, выделяемые в нуклеотидных последовательностях с помощью анализа распределений фрагментов генома. Показано, что последовательность генома хлоропласта обладает невырожденной семикластерной структурой в распределении таких фрагментов по частотам триплетов.

Ключевые слова: порядок, триплет, частотный словарь, фаза, сложность.

Introduction

Molecular biology provides mathematics with a number of mathematically sound problems and questions. Eventually, the structure identification and an order implementation in an ensemble of finite sequences are the most interesting among them. Finite symbol sequences, being a typical mathematical object, are naturally present as genetic matter in any living being; namely, as DNA sequence. Further we will consider the finite symbol sequence of chloroplast genomes of five plant species, including one from *Larix sibirica* Ledeb., which was recently completely sequenced, assembled and annotated in the Laboratory of Forest Genomics at the Genome Research and Education Centre of Siberian Federal University (Krutovsky et al., 2014; Bondar et al., 2015; Sadovsky et al., 2015). This sequence consisted of 122 561 symbols or letters from the four-letter alphabet $\aleph = \{A, C, G, T\}$. Neither other symbols, nor blank spaces are supposed to be found in a sequence; a sequence under consideration is also supposed to be coherent (i.e., consisting of a single piece).

An identification and search of structures in DNA sequence is a main objective of mathematical bioinformatics, biophysics and related scientific fields, including computer programming and information theory. Structures observed within a sequence reveal an order and provide easier

understanding of functional roles of a sequence or its fragments. A new function (or a connection between function and structure, or taxonomy) might be discovered through a search for new patterns in symbol sequences corresponding to DNA molecule.

It is a commonly accepted fact that nucleotide sequences are rather inhomogeneous in terms of a structuredness that is demonstrated in this paper. In particular, any genome sequence roughly comprises two types of subsequences: coding and non-coding ones, respectively. These subsequences usually do not overlap, while their concatenation yields the entire genome (or chromosome) sequence. Basically, the fragments belonging to these two classes differ in their statistical (and/or combinatorial) properties. Some pioneering results in the analysis of this phenomenon see in Gorban et al. (2003, 2005a,b). It was found that the fragments of any genome being converted into special frequency dictionaries demonstrated some specific clusterization in the space of those frequencies. Here we present the results of similar clusterization observed for five chloroplast genomes.

Materials and Methods

Concept

First, we partitioned symbol sequences (that were the chloroplast genomes) for a set of overlapping fragments as long as 303 symbols (nucleotides), starting from the first symbol (nucleotide) at the sequence and then with a shifting window step of 10 symbols (nucleotides) alongside the chloroplast genome sequence. Second, for each fragment in the series described above, a special frequency dictionary was developed. Third, the ensemble of the dictionaries (that was a set of the points in the 63-dimensional Euclidian space) was clustered using the *K-means* technique (Fukunaga, 1990; Mirkes et al., 2013). Forth, the distribution of those fragments over an elastic map is studied (Gorban and Zinovyev, 2009, 2010; Gorban et al., 2008). Finally, a correlation of the fragments belonging to different classes obtained though *K-means* and elastic map implementation to the functionally charged regions of the genome is studied.

Sequence data

The chloroplast genome sequences were used from the following five species: European larch (*L. decidua* Mill.), Norway spruce (*Picea abies* (L.) Karst.) and rice (*Oryza sativa*) obtained from the NCBI GenBank (accession numbers AB501189.1, NC_021456.1 and JN861110.1, respectively), and newly assembled and fully annotated complete chloroplast genome sequences of Siberian larch (*L. sibirica* Ledeb.) and Siberian pine (*Pinus sibirica* Du Tour) that have been sequenced using the Illumina HiSeq2000 sequencer at the Laboratory of Forest Genomics of the Siberian Federal University (Krutovsky et al., 2014; Bondar et al., 2015; Sadovsky et al., 2015).

Lattice and Dictionary

In earlier studies (Bugaenko et al., 1996, 1997, 1998; Hu and Wang, 2001), it was demonstrated that a frequency dictionary can supposedly be used to study fundamental structure of a symbol sequence. Here we introduce the definition of **frequency dictionary.** First, consider a symbol sequence T of the length N from four-letter alphabet. Wherever further we assume that no other symbols (nucleotides) but those from the alphabet $\aleph = \{A, C, G, T\}$ are found in a text $\Im$. Also, no gaps take place in the text. Here we start with a few following definitions:

Definition 1. *The* word $\omega = \nu_1 \nu_2 .. \nu_{q-1} \nu_q$ *of the length* q *is a string occurred in the text* $\Im$. *Here* ν_j *is a symbol occupying the j-th position at the word*; $\nu_j \in \aleph$.

Everywhere below we will consider only the words with the length of 3 symbols (nucleotides) and will call them *triplets*.

Definition 2. (q,l)-frequency dictionary $W(q,l)$ *is the set of all the words of the length* q *counted within the text* $\mathfrak{I}$ *with the step in* l *symbols, so that each word is accompanied with its frequency.*

A frequency of a word ω is defined traditionally: that is the number n_ω of copies of the word divided by the total number of all copies of all the words (Bugaenko et al., 1996, 1997, 1998; Hu and Wang, 2001). Parameter l is arbitrary in a dictionary; everywhere further we will consider only the $W(3,3)$ frequency dictionaries. Evidently, a $W(3,3)$ frequency dictionary comprises a set of codons in some cases.

A frequency dictionary (namely, a dictionary $W(3,3)$) unambiguously maps a text $\mathfrak{I}$ into a 64-dimensional space, where the triplets are coordinate axes in those space, and the frequencies are the coordinate figures. Hence, frequency dictionary represents a short range (or meso-scale, at most) structuredness in a symbol sequence. Consider, then, a frame identifying a fragment F (or a subsequence) of the length S in a text $\mathfrak{I}$.

Definition 3. Lattice *(or* (S,d)*-lattice, to be exact)* $R(S,d)$ *is a set of the fragments consequently identified alongside the text* $\mathfrak{I}$ *by the frame of the length* S*, with the step* d*.* Obviously, a lattice consists of overlapping fragments, if $d < S$.

That is the basic object for further analysis of statistical properties of a symbol sequence representing chloroplast genomes. The key idea of the paper is to check whether the fragments obtained for some (S,d)-lattice $R(S,d)$ differ in their statistical properties, or not. The properties expressed in the terms of (q,l)-frequency dictionaries $W(q,l)$ would be considered, only.

Clusterization techniques

(a) K-means

We used two main techniques to analyze the abundant data: *K-means* and elastic map technique, respectively. *K-means* technique is a good practice in analyzing data of various nature (see for example a classic book by K. Fukunaga (1990)).

In spite of a high popularity, *K-means* has a few problems, which should be taken into account while discussing the results. Stability of a final distribution is the first problem. Since an implementation of a clusterization through *K-means* starts from a random separation of the original data set for K classes, then there is no guarantee the final composition of the classes remains the same. Of course, one might face the situation when the final distribution is identical for any initial separation; this is the situation of the highest stability of clusterization.

On the contrary, there might be a situation where any new start of the procedure yields absolutely other final distribution, and one hardly could recognize any similarity between them. This is the opposite situation of instability; moreover, one should recognize such situation as a total lack of any inner structuredness in the data set. In reality, the situation is somewhere in between. Usually, a set of data splits apparently into two subsets, where the first subset tends to yield rather stable distribution of objects in a series of *K-means* runs, while the other subset gathers the objects that could not be reliably attributed to any class. These are so called volatile objects.

There is no an evident and simple approach to deal with the volatile objects; an elimination of them from the original data set may not guarantee the stability of the classification provided over the rest part of the data set. Evidently, one may consider the method to fail in

classification of this part of the objects. In such a case they should be considered as a ”noise” or background for the stable classification of other subset of data.

Another essential problem of *K-means* is the number of classes determination, as well as the separability of classes. Indeed, there is no an *a priori* way to figure out the exact number of classes for *K-means* classification; usually, that is a matter of expertise of a researcher. This point is related, to some extend, to the previous one: a stability of a classification is not directly related to the number of classes. We did not check stability of the classes and their separability in this research.

(b) Elastic map clusterization

Yet another approach to figure out clusters in a dataset is to implement an elastic map (Gorban and Zinovyev, 2009, 2010; Gorban et al., 2008). The basic idea of this method is to approximate the multidimensional data with a manifold of small dimension; the elastic map technique implies the approximation with two-dimensional manifold (see details in Gorban et al., 2008). In brief, the procedure looks like the following. At the first step, the first and the second principal components must be found. Then a plane must be developed over these two axes. At the second step, each data point must be projected at the plane and connected with the projection by an elastic spring. At the third step, the plane is allowed to bend and expand; so, the system is to be released to reach the minimum of the total energy (deformation plus spring extension). At the fourth step, each data point must be re-determined on the jammed map. Namely, a new data point image is the point on the map that is the closest to the original point in terms of the chosen metrics. Finally, the jammed map is ”smoothened“ by inverse non-linear transformation (for more details see Gorban et al., 2008).

Results

Let now explain the procedure of the chloroplast genome analysis in more detail. Firstly, we covered the symbol sequence with the (S,d)-lattice, with $S = 303$ and $d = 10$; it was important that $d \neq 0 (\mathrm{mod}\ 3)$. Each fragment of the lattice was labeled with the number of symbol occupying the central position of that fragment; that is why S was odd. Next, each identified fragment of the lattice has been transformed into $W(3,3)$ frequency dictionary. Hence, the sequence was mapped into a set of the points in a metric 64-dimensional space. We used Euclidian metrics hereafter.

Actually, the linear constraint

$$\sum_{\omega} f_{\omega} = 1 \qquad (1)$$

brings an additional parasitic signal; therefore, one of the triplets must be eliminated from the set. Indeed, all the points representing various frequency dictionaries are located at the linear subspace of the co-dimension one. Formally, any triplet could be eliminated, but, practically, the choice may affect the results of the further treatment. Here there could be two strategies: either to exclude the triplet that has the greatest frequency or to remove the triplet that made the least contribution into the points separation and discrimination. The first strategy makes sense when the greatest frequency exceeds the others ten times and more. Thus, we have persuaded the second strategy. To do that, the triplet with the minimal standard deviation over the entire dataset was found and excluded; that was the CGC triplet.

So, *K-means* classification was developed, and results are presented in Fig. 1. Evidently, the distribution was not random anyway. It consisted of several distinguishable clusters resembling, to some extend, a bullet, or a shuttlecock.

Regarding the clusterization provided by the elastic map technique to develop the

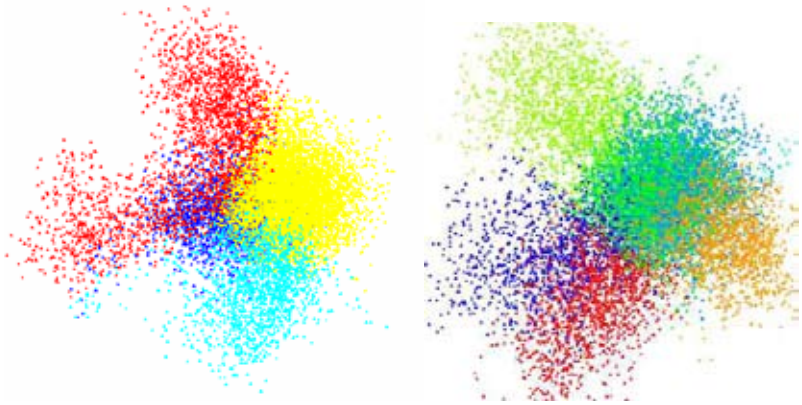

Fig. 1. Distribution of the fragments of (303, 10)-lattice developed for *Larix sibirica* chloroplast genome shown in principal components coordinates. Left: classification for four classes, right: classification for seven classes

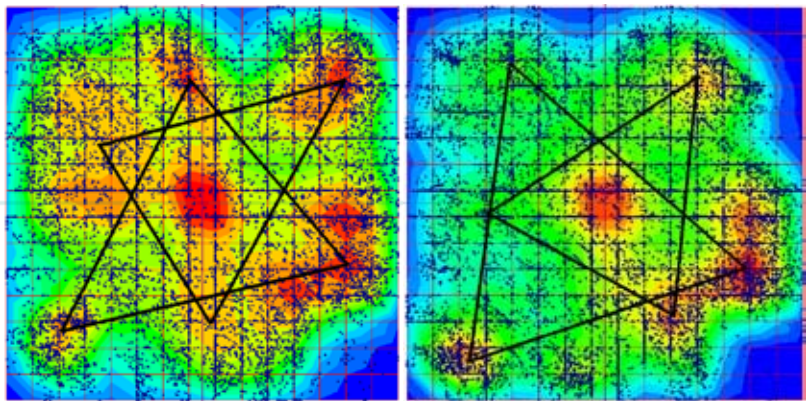

Fig. 2. Elastic map with the fragments of (303, 10)-lattice developed for *Larix sibirica* (left) and *L. decidua* (right) chloroplast genomes shown in inner coordinates. The local density of the fragments on the elastic map is highlighted by colors: red indicates the highest density level, and blue indicates the lowest one

elastic map we used the *ViDaExpert* software by A. Zinovyev and A. Pitenko (http://bioinfo-out.curie.fr/projects/vidaexpert/). The standard parameters configuration was used to develop the map (Fig. 2) that depicts the famous seven cluster structure (Gorban et al., 2003, 2005a,b) more explicitly with due edge-node pattern. The other chloroplast genomes analyzed in this study also showed the seven-cluster structure (Fig. 3). The clusters in Figures 2 and 3 gather the fragments belonging to the functionally different types of the original sequence. Consider, first of all, the central cluster. It consists from the fragments located in non-coding regions of the genome.

Six other clusters comprise two rotated triangles, where the clusters themselves make the triangle vertices. These six clusters correspond to coding regions of the genome. First of all, why triangle? The point is that any sequence may generate three different $W(3,3)$ frequency dictionaries: they differ in the starting position of the first triplet. It is done by selecting the first symbol (nucleotide) in a sequence, and then assigning it as the number one in a fragment. The three different $W(3,3)$ frequency dictionaries

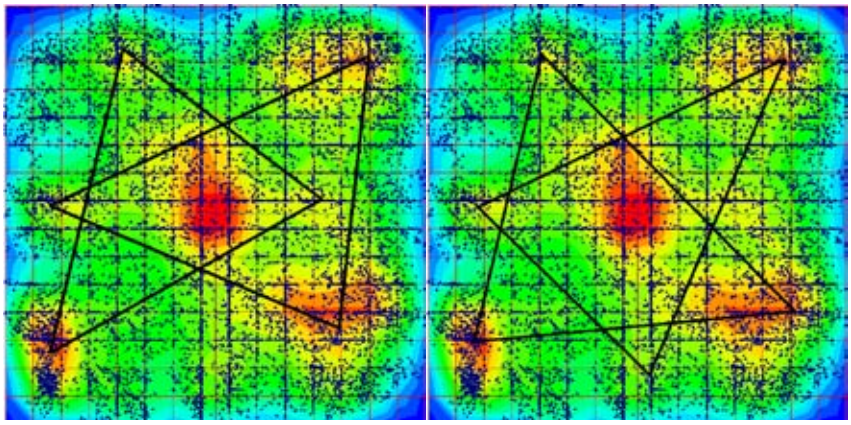

Fig. 3. Seven cluster structure found for *Pinus sibirica* (left) and *Picea abies* (right) chloroplast genomes

could be developed over the fragment: the first one is starting at the given symbol, the second one is starting one symbol apart, and the last one is starting two symbols apart. Thus, these three $W(3,3)$ dictionaries could be labeled as having null-phase, phase 1 and phase 2. So, the vertices of the triangles shown in Fig. 2 correspond to the $W(3,3)$ dictionaries with different phases.

Yet, we did not develop the $W(3,3)$ dictionaries for all three phases within the same fragment; so the question arises where did the dictionaries of different phase come from? The answer is rather evident: the dictionaries of the same phase belong to different coding regions of the genome. More exactly, consider a coding region of the length L to be found somewhere in a genome. The number T_p of $W(3,3)$ frequency dictionaries of the same phase p to be developed within that given region is

$$T_p \approx \frac{L}{3 \times d},$$

where d is the parameter of the (S,d)-lattice. This estimation yields a typical number T_p of $W(3,3)$ dictionaries of the same phase to be found within a single coding region as $T_p \approx 10^2$, and that seems to be the upper limit estimation.

In other words, the coding regions are typically arranged in the genome so that the difference between their starting points k_j^{start} is never divisible by 3:

$$k_j^{\text{start}} - k_l^{\text{start}} \neq 0 (\text{mod}\, 3);$$

here j and l are the numbers of the symbols occupying the starting position of a coding region. It means that an (S,d)-lattice covering a sequence yields an almost homogeneous distribution of the fragments, with respect to the location of a starting position of a coding region observed in a fragment. Two triangles come from two strands of a genome; there could be a degenerated situation where these two triangles (depicted by black lines in Figures 2 and 3) coincide due to a rotation. Such degeneracy takes place, if $n_A \approx n_C \approx n_G \approx n_T$ in a genome. Figures 2 and 3 disprove this relation for the chloroplast genome.

Similar, Fig. 4 demonstrates chloroplast genome structure for rice (*O. sativa*), which was slightly different from other species.

If briefly, the genome exhibits the famous 7-cluster structure (Gorban et al., 2003, 2005a,b). The pattern is not degenerated; moreover, in larch species it has a cluster that seems to be rather questionable from the point of view of its

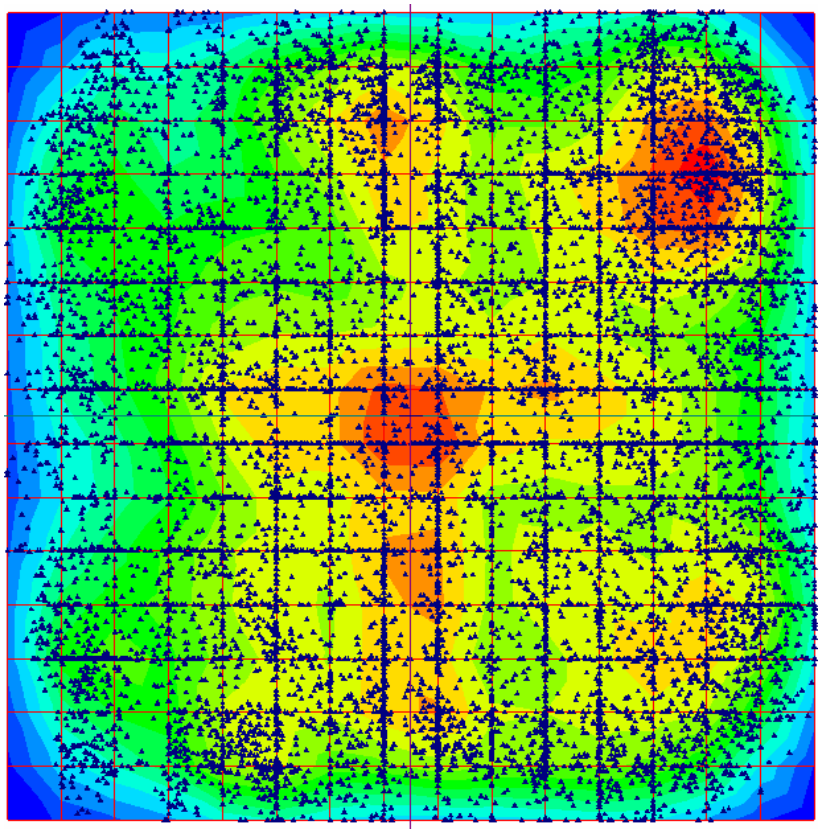

Fig. 4. Seven cluster structure found for rice (*Oryza sativa*) chloroplast genome

fine structure: it looks like a combination of three smaller subclusters (Fig. 2).

Discussion

Here we considered, in some details, statistical structure of chloroplast genomes. A seven cluster pattern is found at the genome. The six clusters arranged in two triangle-like groups correspond to the fragments of the genome occupying the coding regions of that latter. Three different vertices of a triangle correspond to three phases (i.e., to three possible locations of a reading frame against the first symbol of a coding region).

Regarding the dual triangle structure shown in Fig. 2, one may suggest that it resulted from the DNA antisense strand. However, that is not so; we did not develop a dictionary from the opposite strand of DNA. Nonetheless, this dual pattern of three vertices is related to the antisense strand. These clusters consist of the frequency dictionaries representing the so called complimentary palindromic triplets. Two triplets read equally in opposite directions with respect to the Chargaff's complimentary rule make a palindrome. Actually, Chargaff's second parity rule states a close similarity of the frequencies for n_A and n_T, and for n_C and n_G, respectively, counted within a single strand. Generalized Chargaff's second parity rule states the similarity of the frequencies of two strings composing a complimentary palindrome still counted within a strand (Grebnev and Sadovsky, 2014).

Genomes of organisms with various taxonomy ranks differ significantly in terms

of the level of the discrepancy of Chargaff's generalized second parity rule. Mitochondrial DNA leads in terms of a level of violation of the second generalized parity rule. However, in general, a genome in a species at a higher taxonomy rank yields less degree of the discrepancy of the rule (Grebnev and Sadovsky, 2014). Chloroplasts are, in general, the second to mitochondria; yet, the seven cluster structure of the chloroplast genomes may indicate a better execution of the rule for that species. This fact, by the way, may be of great value for a theory of chloroplast origin and evolution: one should seek for the tentative ancestors of chloroplasts a bacterium species that would show similar degree of the second parity rule violation, and exhibit a seven cluster structure of a genome, as well.

Unlike bacterial and yeast genomes (Gorban et al., 2003, 2005a,b), the genomes under consideration demonstrated rather unusual pattern of the clusterization. A split of a vortex of a triangle pattern into the smaller subclusters in a seven cluster structure has not ever been reported. While the chloroplast genomes of *L. decidua* and *L. sibirica* clearly demonstrated this pattern. These are the clusters located at the left and up, and at the right and down in Fig. 2.

The analysis of the fine structure of the seven cluster pattern found at the genomes under consideration may be affected by elimination of a particular triplet. However, to choose the triplet with minimal potential effect to be excluded from the data set, we calculated the standard deviation over the entire dataset; the triplet with the least standard deviation was excluded. This approach seems quite natural and feasible, if the original set of frequency dictionaries is "good" enough. Here it means that a frequency dictionary bears all possible triplets, and their frequencies are different. That should be right for sufficiently long sequences; but the used lattice yielded rather depauperized frequency dictionaries. Therefore, more detailed study is needed to find out the optimal (*S, d*)-lattice for the clusterization search.

Conclusion

Seven cluster structure in chloroplast genomes, including one for *L. sibirica* was found. This is the fundamental structure of any genome; the found pattern is not degenerated since frequency of nucleotide A differed significantly from frequency of nucleotide G. The absence of a degeneracy may indicate the prototypic genome that gave origin to chloroplasts entities; it is supposed to be a bacterial one (following symbiotic theory of organelle origin). Unlike nuclear genome of bacteria, the chloroplast genome yields more complex structure of (at least two) clusters: these seem to consists of two and three subclusters, respectively. The detailed structure of these complex clusters needs more studies, but may bring new understanding of a fine structure details, or of relations between structure and function of chloroplast genome.

Acknowledgements

This study was supported by a research grant № 14.Y26.31.0004 from the Government of the Russian Federation.

References

1. Bondar E. I., Putintseva Yu. A., Oreshkova N. V., Krutovsky K. V. (2015) Study of Siberian larch (*Larix sibirica* Ledeb.) chloroplast genome and development of polymorphic chloroplast markers. In: Proceedings of the 4th International Conference Conservation of Forest Genetic Resources in Siberia, August 24-29, Barnaul, Russia. P. 20-21.

2. Bugaenko N. N., Gorban A. N., Sadovsky M. G. (1996) Towards the definition of information content of nucleotide sequences. Molecular Biology Moscow 30(5): 529-541.
3. Bugaenko N. N., Gorban A. N., Sadovsky M. G. (1997) The information capacity of nucleotide sequences and their fragments. Biophysics 5: 1063-1069.
4. Bugaenko N. N., Gorban A.N., Sadovsky M. G. (1998) Maximum entropy method in analysis of genetic text and measurement of its information content. Open Systems & Information Dyn. 5(2): 265-278.
5. Fukunaga K. (1990) Introduction to statistical pattern recognition. San Diego, London: Academic Press, 578 p.
6. Giancarlo R., Restivo A., Sciortino M. (2007) From first principles to the Burrows and Wheeler transform and beyond, via combinatorial optimization. Theor. Comput. Sci. 387(3): 236-248.
7. Gorban A. N., Zinovyev A. Y. (2009) Principal Graphs and Manifolds. In: Olivas E.S. et al. (eds.) Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques. Hershey, PA, USA: Information Science Reference, IGI Global, p. 28-59.
8. Gorban A. N., Zinovyev A. (2010) Principal manifolds and graphs in practice: from molecular biology to dynamical systems. International Journal of Neural Systems 20(3): 219-232.
9. Gorban A. N., Kögl B., Wunsch D. C., Zinovyev A. Yu. (2008) Principal manifolds for data visuaization and dimension reduction. Lecture Notes in Computational Science & Engineering. V. 58. Springer, Berlin-Heidelberg-New York, p. 124-240.
10. Gorban A. N., Zinovyev A. Yu., Popova T. G. (2003) Seven clusters in genomic triplet distributions. In Silico Biology 3: 39-45.
11. Gorban A. N., Zinovyev A. Yu., Popova T. G. (2005a) Four basic symmetry types in the universal 7-cluster structure of microbial genomic sequences. In Silico Biology 5: 25-37.
12. Gorban A. N., Zinovyev A. Yu., Popova T. G. (2005b) Universal seven-cluster structure of genome fragment distribution: basic symmetry in triplet frequencies. In: Kolchanov N. and Hofestaedt R. (eds.) Bioinformatics of Genome Regulation and Structure II. Springer Science+Business Media, Inc., p. 153-163.
13. Grebnev Ya. V., Sadovsky M. G. (2014) Chargaff's second rule and symmetry in genomes. Fundamental Studies 12(5): 965-958.
14. Hu R., Wang B. (2001) Statistically significant strings are related to regulatory elements in the promoter regions of *Saccharomyces cerevisiae.* Physica A 290: 464-474.
15. Krutovsky K.V., Oreshkova N. V., Putintseva Yu.A., Ibe A. A., Deutsch K. O. Shilkina E. A. (2014) Some preliminary results of a full genome *de novo* sequencing of *(Larix sibirica* Ledeb.) and *(Pinus sibirica* Du Tour.). Siberian Forest Journal 1(4): 79-83 (in Russian, English abstract).
16. Mirkes E. M., Zinovyev A., Gorban, A. N. (2013) Geometrical complexity of data approximators. In: Rojas I., Joya G., and Cabestany J. (eds.) IWANN 2013, Part I, Advances in Computation Intelligence, Springer LNCS 7902, p. 500–509.
17. Sadovsky M. G., Bondar E. I., Putintseva Yu. A., Oreshkova N. V., Vaganov E. A., Krutovsky K. V. (2015) Conifer chloroplast genome has a unique seven cluster structure. Doklady Biochemistry and Biophysics (In press).